Scalable Secure Sharable Storage

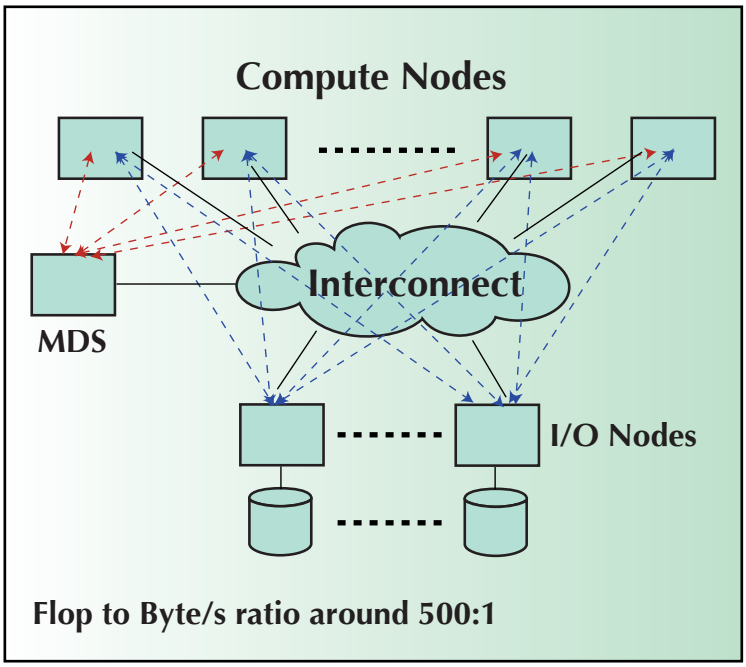

Figure 1. This Parallel I/O architecture offers n *parallel I/O paths to access dedicated storage behind* n *I/O nodes. To allow concurrent accesses of the same file, the parallel filesystem presents a global view to all compute processors via a MetaData Server (MDS): by returning the resolved data location map to its compute clients, the cluster filesystem allows direct I/O operations between compute and I/O nodes in parallel.*

Today's high-speed clusters can easily use the latest interconnect technologies for node-to-node communications, but the I/O storage systems have failed to keep pace. This lag in advancement is of major concern because science, national security, and business computing are becoming more data intensive. Data management challenges will, if not already, exceed the compute-power challenge. Therefore, a versatile storage architecture that is scalable, sharable, and secure is critical to meeting the performance demands for all sectors of High Performance Computing (HPC).

The Scalable Computing R&D Department at Sandia National Laboratories is conducting an R&D project to address the I/O challenges in HPC. Through research and industry partnerships, this study explores Parallel and Remote Direct Memory Access (RDMA) technologies for scalability, promotes standardized common filesystem interface to facilitate sharing, and is compatible with Sandia's corporate security infrastructure.

Parallel Filesystems

Parallel applications move large amounts of data at fixed intervals, between distributed memory and storage, requiring parallel I/O paths to meet HPC demands. Parallel filesystems that allow concurrent accesses and provide parallel paths can greatly ease parallel code development, and significantly simplify post-processing and analysis. Due to the large compute-to-I/O node ratio, large sequential reads and writes from parallel applications become random requests to the backend storage. We are exploring new caching algorithm and RAID configurations that will minimize disk I/O, thereby improving data throughput.

SAND2006-4852P
April 2006

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

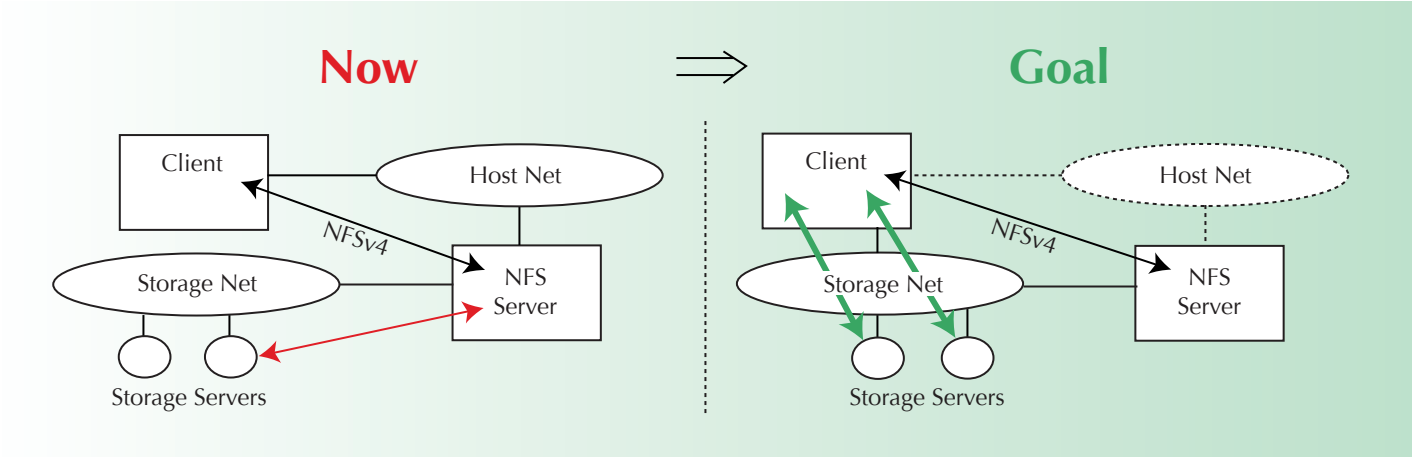

Figure 2. Parallel NFS provides scalable I/O by providing asymmetric, out-of-band control path (open/close) that is different from the parallel data paths (read/write).

Common Filesystem Client Interface

Many open source and commercial parallel filesystems, each require a customized Linux kernel and version. These kernel dependencies create administrative complications and consequently make data sharing between platforms virtually impossible. We have formed strategic alliances with leading storage vendors as well as university research partners to develop a standards-based common client interface, namely Parallel NFS (pNFS), for all advanced filesystems. The pNFS protocol is an extension of NFS v4 that is designed to support parallel storage through the pNFS request routing and/or file virtualization, and the extend block or object "layout" information to the parallel filesystem clients.

Remote Direct Memory Access

Moving data between platforms typically involves multiple copies of the data, from the network interface card to the kernel and then application buffers, on end systems. At 10 Gbps, these copies incur very high CPU overhead and consume large amounts of memory bandwidth, leaving few resources to the processing of scientific applications. This project will integrate emerging RDMA technologies into our storage infrastructure to improve the scalability of HPC I/O subsystems.

Security

Sandia is in the process of upgrading its production security-infrastructure from DCE/DFS to using the LDAP to manage enterprise-wide user accounts, Kerberos5 for cross-realm authentication, and NFS v4 for fine-grained access control of multi-domain data sharing. We are collaborating with our New Mexico counter parts and connected to the NM security testbed for seamless integration.

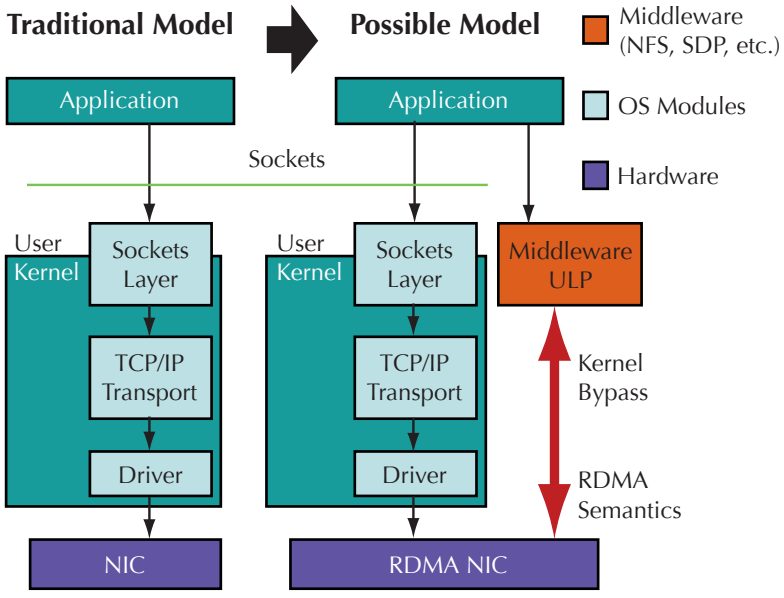

Figure 3. The emerging RDMA technologies can deliver data directly into a remote application's buffer space, thereby eliminating the need for intermediate data copies.

For more information, contact
Helen Chen (925) 294-2991